How text mining algorithms for crowdsourcing can help us to identify today's pressing societal issues

Anna N. Köhl
University of Innsbruck
anna.koehl@student.uibk.ac.at

Simon Fuger
University of Innsbruck
simon.fuger@uibk.ac.at

Moritz N. Lang
University of Innsbruck
moritz.lang@uibk.ac.at

Johann Füller
University of Innsbruck
johann.fueller@uibk.ac.at

Martin R. Stuchtey
University of Innsbruck
martin.stuchtey@uibk.ac.at

Abstract

Crowdsourcing is increasingly applied in the area of open development with the goal to find solutions for today's pressing societal issues. To solve such wicked problems, manifold solutions need to be found and applied. In contrast to this, most recent research in crowdsourcing focuses on the few winning ideas, ignoring the sheer amount of content created by the community. In this study we address this issue by applying an automated text mining technique to analyze the ideas contributed by the crowd in an initiative tackling plastic pollution. We show that automated text mining approaches reveal numerous possibilities to make use of the so far unused content of IT enabled collaboration projects. We further add insights into how our findings can help researchers and practitioners to accelerate the solution process for today's pressing societal issues.

1. Introduction

Our current approach of dealing with natural resources leads to a degradation of natural systems, which builds the invaluable basis of humanity and increases the magnitude of today's biggest societal problems [44, 50]. Complex societal issues are problems with multiple conflicting interests, which over time play against or occasionally with each other [45]. To those kind of wicked problems, there is no single solution, but rather an "*exhaustively describable*" set or series of solutions, e.g., in the form of sustainable innovations [41]. A great number of different stakeholders needs to deal with such problems and change their mindset and behavior to find viable holistic solutions and adequate sustainable innovations. Decentralized, bottom up approaches are required that involve a variety of people with diverse skillsets and different modes of thought [2]. Interesting new applications derive from international development. The emergence of information and communication technologies, the evolution of Web 2.0, and increased internet access in developing countries, enable international development to open up and foster cross-cultural and inter-country collaboration [43]. Named open development, this new model is considered to radically change the development landscape [40] and enables co-creation where the poor are actively engaged in innovation processes [25, 47]. One possibility to include a diverse group of people into problem solving processes and tackle societal issues are crowdsourcing initiatives [13]. With the help of online innovation platforms people are able to share thoughts, work on ideas collaboratively and interact with a broad audience. Using crowdsourcing to exploit collective intelligence has lately received major interest in the field of international development [26]. However, the application of crowdsourcing is not free of problems. Even though crowdsourcing is able to gather a holistic understanding by including a big group of people and perspectives, it has so far mainly been applied in order to identify the single best solutions [5, 22, 30]. This stands in contrast to the characteristics of wicked problems, which are only solvable through a set of solutions [41]. Further, the magnitude of submissions can be overwhelming and idea selection and convergence processes are full of flaws [18, 23]. As an example, Google´s project 10^{100} can be mentioned where over 150.000 submissions were collected which needed an evaluation team of 3.000 internal employees for idea convergence – a resource intensive and doubtful evaluation process [23]. In theory, crowdsourcing is a great source, however we still face the problem of the effort necessary to master the sheer amount of ideas generated [35]. To solve this issue, we apply automated text mining techniques. The technique is applied to the "Circular Design Challenge", a crowdsourcing initiative hosted by the Ellen

URI: https://hdl.handle.net/10125/59509
ISBN: 978-0-9981331-2-6

MacArthur Foundation on openIDEO that tackles the question: "How might we get products to people without generating plastic waste?".

The first research question of this study investigates how researchers and practitioners are able to identify and analyze the crowd generated content. In this sense, this research contributes to existing literature on crowdsourcing by presenting an automated methodology to analyze the content of ideas generated in a crowdsourcing initiative. It also contributes to the current problem that most of the generated content goes unheeded due to the effort to analyze the sheer amount of data. The second research question focuses on getting an overview and a more holistic understanding of the contributed content in the "Circular Design Challenge".

To address these research questions, this study aims to explore the plain content generated within a crowdsourcing initiative focusing on sustainable innovation.

The remainder of the article is structured as follows: In the upcoming section this study introduces literature on sustainable innovation and wicked problems, the research stream of open development linked with the concept of crowdsourcing, and the text mining technique topic modeling. This section is followed by insights on the investigated crowdsourcing community and the applied methodology before the findings of this study are presented. In a last section, this article concludes with a discussion on the outcome of this study.

2. Theoretical background

2.1. Sustainable innovation and wicked problems

The term sustainable innovation has been used increasingly in recent years and includes the holistic and long-term development of social, economic and ecological objectives [6, 10]. Such innovations are impact oriented, systemic and transformational and follow a boundary breaking and inclusive approach [6, 10, 24]. This type of innovation can be seen as relevant for solving wicked problems [24].

A typical wicked problem is environmental pollution, and plastic pollution in specific, as it has innumerable causes, is tough to describe in its complete magnitude and there is no easy way out [41, 49]. Levin et al. [31] even go one step further describing climate change (including pollution) as a super wicked problem which additionally comprises the four key features: *"[…] time is running out; those who cause the problem also seek to provide a solution; the central authority needed to address them is weak or non-existent; and irrational discounting occurs that pushes responses into the future (p.124)."* To such super wicked problems, there is no single solution, but rather an *"exhaustively describable"* set or series of solutions [41]. As plastic pollution has become a pressing issue, science and practice has increasingly focused on finding viable solutions [32, 54]. The discussion shifted from effect-oriented solutions to source-oriented solutions [32]. Effect-oriented approaches only adapt to the consequences [32], whereas the latter are solutions that stop and/or minimize pollution in the first place by changing production and consumption [48]. The solution space includes the three generic forms of product innovation, technology innovation, and behavioral and system innovations. Product innovation is the introduction of a good or service that is new or significantly improved with respect to its characteristics or intended uses [38], whereas technology innovation focuses on the invention or significant change of the underlying technology [19, 37]. Sustainable product and technology innovations go beyond the improvement of its characteristics or intended uses, as they also include the reduction of resources to meet global societal and natural pressures [14]. As sustainable innovation goes beyond product and technology innovation [24], behavioral and system changes are also included. System innovation comprises behavioral changes as it is defined as transformation in the way societal functions like transportation, shopping and feeding are fulfilled [16]. The solution space for plastic packaging therefore has to consist of product, technology and system innovation.

Nevertheless, as it is a wicked problem, many solutions do have other problems as a consequence [41]. For example the technology innovation of biodegradable plastic has the undesirable side effect of negatively influencing the recycling stream of conventional plastics [53]. It becomes obvious that the solution space today is not exhaustive and that new ideas need to be generated. A great number of different stakeholders from all over the world need to deal with the issue and change their mindset and behavior. As the term sustainable innovation also encompasses firm external forms of innovation, e.g., open innovation, sustainable innovation democrats solution finding as it aims at including all people [24]. This is in line with the evolution on the international development, where open approaches are increasingly focused.

2.2. Open development as a model to foster sustainable innovation

Recently, researchers suggest that "open" models allow for increasing the effectiveness of developing innovations that address the roots of a societal problem, like for example plastic pollution [1, 9]. Simultaneously, researchers argue that tools focusing on international development are becoming more "open". Processes supported with information and communication technologies (ICT) that bring people together and improve their lives are in the rise [1, 40]. Named open development, this new model is attributed to the ability to radically change the development landscape [40].

According to Heeks [25] and Thompson [47] ICT enables bottom-up collaboration and co-creation where the poor are actively engaged in innovation processes. Due to the evolution of ICT and Web 2.0, cross-cultural and inter-country collaboration becomes possible and enables the share of ideas and the reuse and revision of content [43]. Cheng et al. [11] highlight the hurdles of multi-cultural collaboration and state that multinational cooperation needs to be supervised to build stable trust among participants [12]. In addition, the transparency of processes can be increased with the help of ICT [43].

One popular possibility of implementing open development projects with the help of ICT is the concept of crowdsourcing. Crowdsourcing is an approach that makes use of the wisdom of the crowd, meaning to take a job traditionally performed in-house and outsource it to a bigger group of individuals, known as the crowd [27]. Existing research has shown the potential of using the concept of crowdsourcing to solve pressing societal issues [8, 34]. Also Nielsen [36] argues that the participants of open development projects in the field of sustainable innovation are important innovating actors. Since Howe [46] introduced the term crowdsourcing in his Wired Magazine article, clearly a lot has happened. Significant technological developments and the ease of using information technology enabled the mass to participate in different kinds of crowdsourcing initiatives and the potential future applications are manifold [29]. Most recently, Kietzmann [29] suggests a new and revised definition of the concept crowdsourcing: *"The use of IT to outsource any organizational function to a strategically defined population of human and non-human actors in the form of an open call (p.2)."*

Crowdsourcing is commonly used by companies to outsource a specific idea generation process. This idea generation process can also be used by public organizations that search for ideas to solve certain societal issues. Typically, an idea contest starts with a problem statement presented by the initiator followed by an idea phase where members of the crowd are able to post their ideas based on the given problem statement. At the end of such a crowdsourcing initiative winning ideas are selected by experts (or a community voting) and rewarded by either a monetary price or support in realizing the idea.

Even though crowdsourcing is able to gather a holistic understanding through the collective intelligence, it has so far mainly been applied in order to identify the best solutions, losing the content of all other ideas [5, 22, 30]. Researchers argue that this idea evaluation and selection is best done by including a convergence phase where a collection of ideas is reduced and clarified [18, 23, 42]. Literature states that this idea convergence phase can be more time consuming than the idea generation itself [15, 20]. While many ideas are generated, such contests are until now not analyzed in order to get a holistic understanding about the topic, as the effort is too big and is not paying off. However, we argue that in the area of societal problems and crowdsourcing for open development, it is essential to get a holistic understanding about the solution space and the problem. As crowdsourcing initiatives deliver rich data-sets, which is one of the main requirements for data mining, we apply text mining to profit from the data ubiquity of crowdsourcing in order to get a holistic understanding of complex problems.

2.3. Text mining through topic modeling

Text mining is the process of deriving high-quality information and hidden relations from text. It is a relatively novel approach, rooted in the idea of data mining [33, 52]. A frequently used tool for text mining is topic modeling, which exploits the correlations among the words and latent semantic themes [3].

"Topics" stand for the hidden (latent) structure of the text that link specific sets of words to their occurrence in documents. Each text is not only consisting of one topic, but can simultaneously consist of several topics, with varying topic distributions over the document [21]. The derived topics explain similarities between different texts and reveal hidden structures and relations between content.

Topic modeling is an automated analysis with no interference or interpretation by the researcher. It has already been applied in different research areas, as it is viable to be used for domain specific purposes [51] and can help to understand large amounts of unstructured text and data. It offers high-quality clustering and detects hidden structures. As a research method, topic modeling has the advantages of handling large amounts

of data in a short time period. Text mining has received increasing attention and many researchers see it as a cure to several issues. However, it requires certain conditions in order to generate meaningful outcomes, e.g., a big data set. Thus, it may work well in combination with crowdsourcing where we generate a ubiquitous amount of data which is difficult to handle.

In this research we apply the Latent Dirichlet Allocation (LDA) algorithm [4], one of the most prominent topic models [51], to investigate whether automated text mining helps to make sense of the otherwise unused content and to get a better understanding of the solution space in the area of plastic pollution.

3. Methodology

3.1. Case description

OpenIDEO is a platform that hosts a community of more than 17 000 users from over 170 different countries. OpenIDEO only provides the platform and the community, acts as a facilitator, and hosts challenges from different initiators ranging from governmental organizations and NGO´s, to private companies. The community at openIDEO can be considered as a collaborative community [7].

The initiative of interest is hosted by the Ellen Mac Arthur Foundation, called "Circular Design Challenge", and tackles the question: "How might we get products to people without generating plastic waste?". Participation was open to everybody, with monetary incentives for winning ideas. Contributions were evaluated by a jury of experts. The challenge was launched in May 2017. Over a period of 6 month, participants were asked to submit ideas. Top ideas were announced in October 2017. Idea creators of winning ideas received monetary support to execute and implement the elaborated ideas.

3.2. Data

Data was scraped from the platform openIDEO in October 2017 including all public data about participating users, data about each submitted idea including text, and data about each and every user interaction in form of comments throughout the challenge. In total, 1107 users participated in the "Circular Design Challenge" meaning they took part in at least one interaction with another user. 72% of all users stated their country of origin, implying that data about country of origin is present from 801 individuals. Figure 1 presents the geographic allocation of those 801 participants.

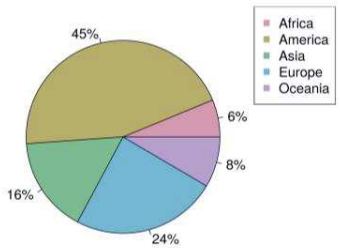

Figure 1. Geographic allocation of participants

Overall 619 ideas were submitted by 483 individuals. 100 ideas were evaluated to be most promising by the expert jury and idea creators received the chance to further work on their idea in the refinement phase. At the end of the challenge 16 ideas were rewarded as winning ideas. Figure 2 shows the country of origin of idea creators within the different phases.

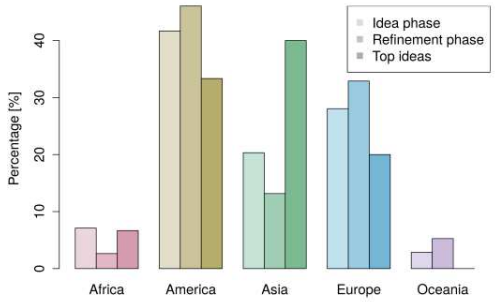

Figure 2. Idea creators' country of origin

The word count of all ideas together equals 219 244 words which translates to 381 pages in a single document.

3.3. Method

Given the sheer amount of text we propose to use automated text mining techniques to analyze the content of ideas. As described in section 2.3., topic modeling exploits the correlations among the words and latent semantic themes [3]. As the LDA is one of the most commonly used algorithms for text mining [51], we use it as a tool to analyze our dataset of 619 ideas and 219 244 words. Through the help of LDA, we are able to get an understanding of the content of the ideas, without reading the 381 pages full of text [35]. The LDA generates topics that cluster the content of the ideas, reveal the hidden relations between the topics and explains similarities between different ideas. These outcomes of the LDA correspond to the context of our research and help to answer our research questions.

The LDA is a Bayesian mixture model assuming that word and topic probabilities follow a Dirichlet distribution. It generates coherent topics through an iterative process of assigning words to a contemporary topic and then repeatedly checking and updating this temporary topic assignment. The process of checking the topic assignment is cycling through the entire collection of text, in our case ideas, multiple times to assign every word to one or multiple topics and to define every text by a range of different topic probabilities. The iterative process of the algorithm is implemented using a technique called Gibbs sampling and checks how prevalent every word is across topics, and how prevalent every topic is in each idea.

In the following, the preparation of data and application of topic modeling on the openIDEO dataset will be described. The data preparation is performed by the package *tm* [21], which provides an excellent text mining infrastructure for the programming language *R* [17]. The fitting is performed using the *topicmodels* package applying a Gibbs sampling algorithm [21].

All ideas with a missing idea text (incomplete ideas with no idea description given) were omitted from the dataset (n=23) resulting in 596 ideas relevant for analysis. In a first step, all idea texts were transformed into a single string of text, called corpus. The corpus is then processed to remove all stop words (e.g., "a", "by", "the"), punctuations, and non-standard UTF-8 characters. Additionally, the terms are stemmed, meaning they are reduced to their word stem. In our research context focus is on solutions, therefore it is important that the words of the problem formulation are not biasing the analysis. To minimize this bias, we made use of the mean term frequency-inverse document frequency (tf-idf). Tf-idf allows omitting terms which have low frequency as well as terms occurring in most documents. We only include terms which have a tf-idf value greater than the median, which ensures that the very frequent terms which have no influence on a specific idea content because they appear in every idea as they are part of the problem formulation (e.g., "plastic") are omitted.

Figure 3. Most frequent words within the corpus after tf-idf selection

Figure 3 shows the word cloud with most frequent words within the corpus after performing the tf-idf selection.

For fitting the LDA model a predefined number of topics must be determined. As the optimum number of topics is not known, the out-of-sample log-likelihood is analyzed for a different number of topic models ranging from 2 to 100. For each the data is divided into 10 different subsets, and each subset gets one turn as validation set and 9 turns as part of the training set. Figure 4 shows that the mean out-of-sample log-likelihood (blue line) has a maximum at 34 topics. But as illustrated by the red colored shading, the number of topics between 30 and 45 have only a minor impact on the out-of-sample performance. To keep the interpretation of the topics as simple as possible without any performance loss, we choose 30 topics for further analysis. We finally fit an LDA model with 30 topics performing a Gibbs sampling with a burn-in of 1000 iterations and recording every 50th iteration for 2000 iterations. By default only the best model with respect to the posterior likelihood observed during Gibbs sampling is returned.

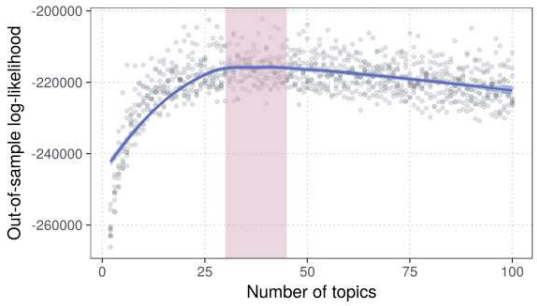

Figure 4. Optimum number of topics

To allocate topics to ideas the approach of set memberships is taken, meaning that each idea essentially is a set of its constituent words. This implicates that certain topics appear with a certain percentage in a specific idea. An example is provided in table 1 below.

Table 1. Idea 160: composition

	Topic 5	Topic 18	Topic 23	Topic 4	Topic11	…
Idea 160	81.9%	2.4%	1.9%	1.4%	1.4%	…

The topics derived from the performed automated LDA analysis have important characteristics. Two ideas may have some topics in common as they appear in both ideas. Nevertheless, the importance of a topic in an idea is defined with its frequency, meaning a topic may appear in more than one idea but differs in probability.

4. Results

4.1. Topics

In table 2 we present the topics derived from all idea texts from the openIDEO dataset. Each topic is described by 4 words, but note that a topic consists of many more words with decreasing probabilities. Also note that words are stemmed as described before. To give the topics a semantic meaning, we defined indicative qualitative labels for the topics. We used the three most frequent words in a topic to derive the qualitative labels (topic names). It is important to understand that these labels do not influence the algorithm or any further analysis.

Table 2. Topics

Nr	Topic Name	Word 1	Word 2	Word 3	Word 4
01	"Opening"	film	seal	top	attach
02	"Water"	ocean	pollut	edibl	sea
03	"Reutilization"	reusabl	item	custom	return
04	"Bulk"	contain	deliveri	bulk	zero
05	"Influenceable degradation"	bottl	fibr	coat	barrier
06	"Dissolvable packaging"	pack	stick	beverag	sugar
07	"Waste"	build	hous	chemic	school
08	"Lid"	bottl	cap	pet	adapt
09	"Lunch"	offic	kid	snack	utensil
10	"Squeezer"	box	juic	vessel	custom
11	"Deposit machine"	machin	deposit	supermarket	item
12	"Education"	suppli	zero	network	educ
13	"Gamification"	app	eco	bin	reus
14	"Monetarization"	prize	lotteri	print	kiosk
15	"Refill sachets"	sachet	shampoo	dispens	refil
16	"Natural material"	paper	hemp	plant	fiber
17	"Island"	inform	puerto	rico	format
18	"Biodegradable"	biodegrad	compost	pla	replac
19	"Drink"	water	drink	bottl	fountain
20	"Straws"	straw	plant	nsheke	bamboo
21	"Cosmetics"	liquid	soap	altern	travel
22	"Animalistic decomposition"	mealworm	chicken	eat	styrofoam
23	"Refill station"	refil	pump	simpl	activ
24	"Package free"	custom	bag	shop	home
25	"Coffee"	coffe	milk	fresh	ship
26	"Toothpaste"	tube	layer	toothpast	inner
27	"Dishes"	leav	wrap	banana	cloth
28	"Convenience"	sauc	restaur	hotel	shape
29	"Multi-functional"	item	label	format	separ
30	"Coffee lid"	cup	lid	coffe	drink

To visualize the derived topics and their connections we construct a network of word distributions over topics (figure 5).

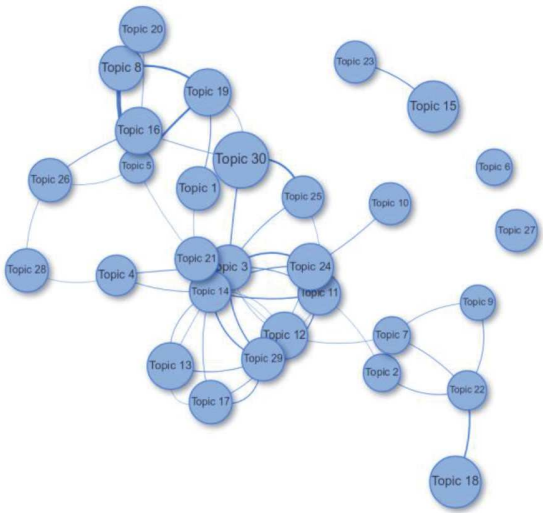

Figure 5. Topic network

The network is created by using the correlation between each topic's word probabilities. The node size reflects how often the topic occurs in the ideas. The lines show the correlation between the topics according to their overlapping word probabilities. The proximity of the topics reflects how interdependent the topics are, again depending on their overlapping word probabilities. The network is done using the R package *LDAvis* and all edges with correlation lower that 5% are omitted. One can see that for example topic 6 and topic 27 are completely isolated and do not have any connection to any other idea, and topic 15 and 23 are only connected to each other. In contrast, topic 14 is very central having direct connections to 9 other topics. Topic 30 is the biggest topic and topic 5 the smallest topic due to their occurrence probability.

4.2. Differences of idea compositions

In a next step, this study elaborates differences in the composition of ideas. An example of an idea composition was already presented in table 1, showing an idea and its' distribution of contained topics. We test whether there are differences in the idea composition between all ideas and successful ideas, successful being defined as reaching the refinement phase. Two groups of ideas are compared with respect to the idea composition: The ideas that were selected into refinement phase (n=99) against the ideas that were not selected to the refinement phase (n=497). We ran an independent sample T-test for all 30 topics, and significant differences ($p<0.05$) in mean values between these two groups are present in ten topics (Table 3).

Table 3. Topic differences

Topic	Topic Name	Delta Δ
03	"Reutilization"	+2.81%*
07	"Waste"	-1.26%***
11	"Deposit machine"	-0.91%*
14	"Monetarization"	-0.98%*
15	"Refill sachets"	+2.50%*
16	"Natural Materials"	-1.63%***
17	"Island"	-1.02%*
18	"Biodegradable"	-1.74%***
22	"Animalistic decomposition"	-1.01%*
27	"Dishes"	-0.99%*

Note: *p<0.05; **p<0.01; ***p<0.001

Taking topic 18 "Biodegradable" as an example: Ideas that did not reach the refinement phase on average consist to 4.13% of topic 18. In contrast ideas that did reach the refinement phase on average consist to 2.38% of topic 18. This results in a significant mean difference of -1.74% (p<0.001).

As discussed in chapter 3.3., each idea consists of several topics with respective probabilities. The topic with the highest probability per idea is referred to as the top topic. We test whether there is a significant difference in the mean probabilities of the top topics for ideas reaching and not reaching the refinement phase by an independent sample T-test. The mean probabilities of the main topics for ideas not reaching the refinement phase is 18.32% compared to ideas reaching the refinement phase with 26.96%. In other words, the top topic in an idea that did reach the refinement phase is on average 8.64% (p<0.001) more prominent than the top topic in an idea that did not reach the refinement phase.

4.3. Qualitative clustering of topic affiliations

In addition to these quantitative analyses we conduct a qualitative clustering based on the theory of sustainable innovation presented in chapter 2.1.

Table 4. Qualitative clustering

	Cluster	Topics
Source-oriented	Product Innovation	Opening (1), Lid (8), Squeezer (10), Deposit machine (11), Straws (20), Cosmetics (21), Toothpaste (26), Dishes (27), Convenience (28), Coffee lid (30)
	Technology Innovation	Influenceable degradation (5), dissolvable packaging (6), Natural Material (16), Biodegradable (18), Animalistic decomposition (22), Multi-functional (29)
	Behavioral and system change	Reutilization (3), Bulk (4), Education (12), Gamification (13), Monetarization (14), Refill Sachets (15), Refill station (23), Package free (24)
Effect-oriented		Water (2), Waste (7), Lunch (9), Island (17), Drink (19), Coffee (25)

Based on the first 10 words of every topic, we qualitatively grouped the topics in the following four clusters: Product innovation, technology innovation, behavioral and system change, and effect-solutions (Table 4). We assess two randomly chosen ideas for each of the 30 topics, only addressing top topics. For example on topic 1 "Opening" only ideas are considered where the topic 1 is the top topic.

As an example we present the idea with ID 487 "PushPop Bottle", which has topic 1 "Opening" as top topic with 7.5%. The idea is a single bottle with integrated push-to-drink closing mechanism removing the need for multiple plastics and detachable bottle caps. The novel product (packaging) is described in detail, see figure 6. Based on the definition of product innovation we can verify that in this case the topic 1 "Opening" can be clustered in the category product innovation.

Figure 6. Idea 487: PushPop Bottle

5. Discussion and implications

On the example of an open development project focusing on sustainable innovation, we examine how researchers are able to analyze the crowd generated content with the help of topic modeling and try to get an overview and a more holistic understanding of the contributed content of the "Circular Design Challenge". The investigated initiative provides valuable proposals to fight plastic pollution and an LDA algorithm enables us to grasp valuable insights on the crowd generated content and allows researchers and practitioners to make use of these insights.

Despite the growing accessibility of ICT and Web 2.0 in developing countries, only 27% of all idea creators come from developing countries. Nevertheless, findings reveal that 47% of winning ideas were contributed by individuals from developing countries. The effects of plastic pollution are especially perceptible in developing countries [28], and the high percentage of winning ideas from developing countries shows, that ICT enabled open development can help to

integrate ideas and solutions of affected individuals and fosters bottom-up collaboration and inter-country collaboration [2, 25, 43, 47]. Due to their direct contact to the effects of pollution [28], individuals from developing countries may address such problems with a more source-oriented perspective than others, which implicates that it is important to integrate those affected individuals [43].

Due to the huge amount of resources needed to manually analyze the sheer amount of generated content within crowdsourcing initiatives, focus is mostly set on the winning idea(s) and ignores all other ideas [5, 22, 30]. With the help of automated text mining, we are able to analyze 596 ideas, equivalent to 219 244 words or 381 pages, into 30 topics including their relations to each other and their importance. We show that the LDA as a text mining tool is a very powerful instrument for crowdsourcing initiatives, as it opens up completely new opportunities to make use of the generated content. In this sense, huge amount of data can be handled in a relatively short time period. Also, this methodology can be of interest for research on convergence in collaborative communities. The LDA algorithm not only allows investigating the relation of selected ideas in terms of content, but it is also capable to simplify the assessment of convergence quality described by Seeber et al. [42].

Through the help of the automated text mining, we are able to show that the occurrence of certain topics affects the performance of an idea. The analysis shows that an idea that includes topic 2 “Reutilization” and topic 15 “Refill Sachets” has a higher probability to enter the refinement phase than other ideas. These two topics are both behavioral changes, which is in line with the definition of sustainable innovation stating that such innovations need to go beyond simple product innovation [6]. In contrast, certain topics have a negative correlation with idea success, meaning that their presence decreases the likelihood of being successful. Such topics are for example topic 7 “Waste” and topic 17 “Island “, both effect-oriented solutions. Another finding of this study is the importance of focus in an idea. We show that an idea with a strong focus on one specific topic is more likely to be successful than an idea evenly composed of several topics.

As discussed in the analysis of the generated ideas, text mining can contribute to a better understanding of the problem and solution space of the tackled issue. The findings reveal the occurrence probability of the topics within the initiative and show which topics are of special importance for the crowd (Figure 5). Through collating this with the current solution space, areas in need for further focus can be identified. For example topic 30 is very prominent, but many initiatives worldwide already exist focusing on how the waste of coffee-to-go cups can be eliminated [39]. Researchers and practitioners can extrapolate that further focus on coffee-to-go cups may not be necessary, as the crowd already focuses on it. On the contrary, topic 5 “influenceable degradation” has the lowest relevance. Through focus on ideas within that topic, researchers and practitioners can investigate the potential of these ideas and examine if further acceleration and support is needed. This elaboration clearly shows the possibilities researchers and practitioners have by using text mining methods. Further research on the specific topic of plastic pollution can be conducted by further analyzing the findings presented in this study.

Our results have important theoretical and practical implications, as they show that the understanding of content of IT enabled open development initiatives can be of high interest. For researchers in the area of crowdsourcing we unveil a method that can be an opportunity to ease the process of understanding and analyzing the crowd generated content of such initiatives. The presented findings demonstrate that practitioners of open development can make use of the concept of crowdsourcing to include people from developing countries into idea generation processes. The ideas of users from developing countries are very promising, as they meet pressing needs not only in their environment, but also in the rest of the world. The results can aid institutions to make better use of IT enabled open development projects. The identification of topics addressed by the crowd can help practitioners to identify the most relevant and pressing needs on a given societal issue. The understanding of the current solution space can help to accelerate the search for solutions to pressing societal issues.

6. Conclusion

There are many societal issues to which there is no single solution. In today´s society human resources, time and money are scarce. Automated solutions are needed to make better use of content generated by the crowd in open development projects.

This study shows why IT enabled collaboration techniques are extremely relevant for development, but also how this valuable content that is generated can be analyzed. We apply an automated text mining technique that is capable of analyzing a vast amount of individual ideas in a crowdsourcing initiative. This methodology needs to be applied to additional datasets of different crowdsourcing initiatives to verify the effectiveness and meaningfulness of topic generation. With the help of the LDA algorithm we reveal the solution space generated by several hundred

individuals on the wicked problem of plastic pollution. Developing countries are most affected by such societal issues. These countries are first profiteers of solutions to these problems. With the help of the presented methodology a better holistic understanding about an open development project can be gained, leading to a better understanding of problems in developing countries. With this information better solutions can be implemented in such regions. This study shows a new way of analyzing crowdsourced data and invites researchers to make use of open development models and the presented text mining technique to gain a better understanding on the solving of societal issues.

7. Acknowledgements

This research was undertaken as part of Anna Köhl's participation in the Schmidt MacArthur Fellowship.

8. References

[1] Bentley, C.M. and A. Chib, "The Impact of Open Development Initiatives in Lower-and-middle-income countries: A Review of the Literature", The Electronic Journal of Information Systems in Developing Countries(74), 2016, pp. 1–20.

[2] Bisgaard, T. and C. Høgenhaven, Creating new concepts, products and services with user driven innovation. Nordic Council of Ministers., FORA, 2010.

[3] Blei, D. and J. Lafferty, "A correlated topic model of Science", Annals of Applied Statistics, 1(1), 2007, pp. 17–35.

[4] Blei, D., A. Ng, and M.I. Jordan, "Latent Dirichlet allocation", Journal of Machine Learning Research, 3, 2003, pp. 999–1022.

[5] Blohm, I., C. Riedl, J. Füller, and J.M. Leimeister, "Rate or trade? Identifying winning ideas in open idea sourcing", Information Systems Research, 27(1), 2016, pp. 27–48.

[6] Boons, F., C. Montalvo, J. Quist, and M. Wagner, "Sustainable innovation, business models and economic performance: An overview", Journal of Cleaner Production, 45, 2013, pp. 1–8.

[7] Boudreau, K.J. and K.R. Lakhani, "Using the crowd as an innovation partner", Harvard Business Review, 91(4), 2013, pp. 60–69.

[8] Brunswicker, S., V. Bilgram, and J. Fueller, "Taming wicked civic challenges with an innovative crowd", Business Horizons, 60(2), 2017, pp. 167–177.

[9] Chalmers, D., "Social Innovation: An exploration of the barriers faced by innovating organisations in the social economy", Local Economy, 2012, pp. 1–18.

[10] Charter, M., C. Gray, T. Clark, and T. Woolman, "Review: The role of business in realising sustainable consumption and production", in Perspectives on radical changes to sustainable consumption and production: 1. System innovation for sustainability, A. Tukker, M. Charter, C. Vezzoli, E. Stø, and M.M. Andersen, Editors. 2008. Greenleaf: Sheffield, England.

[11] Cheng, X., S. Fu, and D. Druckenmiller, "Trust Development in Globally Distributed Collaboration: A Case of U.S. and Chinese Mixed Teams", Journal of Management Information Systems, 33(4), 2017, pp. 978–1007.

[12] Cheng, X., S. Fu, J. Sun, Y. Han, J. Shen, and A. Zarifis, "Investigating individual trust in semi-virtual collaboration of multicultural and unicultural teams", Computers in Human Behavior, 62, 2016, pp. 267–276.

[13] Christensen, C.M., H. Baumann, R. Ruggles, and T.M. Stadtler, "Disruptive innovation for social change", Harvard Business Review, 84(12), 2006, pp. 1–8.

[14] Dangelico, R.M., D. Pujari, and P. Pontrandolfo, "Green Product Innovation in Manufacturing Firms: A Sustainability-Oriented Dynamic Capability Perspective", Business Strategy and the Environment, 26(4), 2017, pp. 490–506.

[15] den Hengst, M. and M. Adkins, "Which collaboration patterns are most challenging: A global survey of facilitators.", 40th HICSS'07, Big Island, Hawaii, 2007, 17b.

[16] Elzen, B., F.W. Geels, and K. Green, eds., System innovation and the transition to sustainability: Theory, evidence and policy, Edward Elgar, Cheltenham, UK, Northhampton, MA, USA, 2004.

[17] Feinerer, I., K. Hornik, and D. Meyer, "Text Mining Infrastructure in R", Journal of Statistical Software, 25(5), 2008.

[18] Fu, S., G.J. de Vreede, X. Cheng, I. Seeber, R. Maier, and B. Weber, "Convergence of Crowdsourcing Ideas: A Cognitive Load Perspective.", Proceedings of the 34rd International Conference of Information System (ICIS), Seoul, Korea, 2017.

[19] Garcia, R., "A critical look at technological innovation typology and innovativeness terminology: A literature review", Journal of Product Innovation Management, 19(2), 2002, pp. 110–132.

[20] Girotra, K., C. Terwiesch, and K.T. Ulrich, "Idea Generation and the Quality of the Best Idea", Management Science, 56(4), 2010, pp. 591–605.

[21] Grün, B. and K. Hornik, "topicmodels: an R package for fitting topic models", Journal of Statistical Software, 13(40), 2011.

[22] Guth, K.L. and D.C. Brabham, "Finding the diamond in the rough: Exploring communication and platform in crowdsourcing performance", Communication Monographs, 84(4), 2017, pp. 510–533.

[23] Haller, J.B., K. Hutter, J. Füller, and K.M. Möslein, "Play or Vote: Matching Games as New Approach for Design Evaluation in Innovation Contests.", in Handbook of research on serious games as educational, business and

research tools, M.M. Cruz-Cunha, Editor. 2012. IGI Global: Hershey, Pa.

[24] Hautamäki, A. and K. Oksanen, "Sustainable Innovation: Solving Wicked Problems through Innovation", in Open innovation: A multifaceted perspective, A.-L. Mention and M. Torkkeli, Editors. 2016. WORLD SCIENTIFIC: Singapore.

[25] Heeks, R., "ICT4D 2.0: The Next Phase of Applying ICT for International Development", Computer, 41(6), 2008, pp. 26–33.

[26] Hellström, J., "Crowdsourcing Development: From Funding to Reporting", in The Palgrave Handbook of International Development, J. Grugel and D. Hammett, Editors. 2016. Palgrave Macmillan UK: London, p.l.

[27] Howe, J., Crowdsourcing: Why the power of the crowd is driving the future of business, 1st edn., Crown Business, New York NY, 2009.

[28] Jambeck, J.R., R. Geyer, C. Wilcox, T.R. Siegler, M. Perryman, A. Andrady, R. Narayan, and K.L. Law, "Marine pollution. Plastic waste inputs from land into the ocean", Science (New York, N.Y.), 347(6223), 2015, pp. 768–771.

[29] Kietzmann, J.H., "Crowdsourcing: A revised definition and an introduction to new research.", Business Horizons, 60(2), 2017, pp. 151–153.

[30] King, A. and K.R. Lakhani, "Using open innovation to identify the best ideas", MIT Sloan Management Review, 55(1), 2013, p. 41.

[31] Levin, K., B. Cashore, S. Bernstein, and G. Auld, "Overcoming the tragedy of super wicked problems: Constraining our future selves to ameliorate global climate change", Policy Sciences, 45(2), 2012, pp. 123–152.

[32] Löhr, A., H. Savelli, R. Beunen, M. Kalz, A. Ragas, and F. van Belleghem, "Solutions for global marine litter pollution", Current Opinion in Environmental Sustainability, 28, 2017, pp. 90–99.

[33] Mehler, A. and C. Wolff, "Einleitung: Perspektiven und Positionen des Text Mining", LDV-Forum, 20(1), 2005, pp. 1–18.

[34] Michelucci, P. and J.L. Dickinson, "HUMAN COMPUTATION. The power of crowds", Science (New York, N.Y.), 351(6268), 2016, pp. 32–33.

[35] Neuendorf, K.A., The content analysis guidebook, 9th edn., SAGE Publ, Thousand Oaks, 2010.

[36] Nielsen, K.R., Crowdfunding for Sustainability: A study on the potential of reward-based crowdfunding in supporting sustainable entrepreneurship, Frederiksberg, 2017.

[37] OECD, The nature of innovation and the evolution of the productive system. technology and productivity-the challenge for economic policy, 1991.

[38] OECD, The Measurement of Scientific and Technological Activities: Guidelines for Collecting and Interpreting Innovation Data, 2005.

[39] Poortinga, W. and L. Whitaker, "Promoting the Use of Reusable Coffee Cups through Environmental Messaging, the Provision of Alternatives and Financial Incentives", Sustainability, 10(3), 2018, p. 873.

[40] Reilly, K. and M.L. Smith, "The emergence of open development in a network society", Open development: Networked innovations in international development, 2013, pp. 15–50.

[41] Rittel, H.W.J. and M.M. Webber, "Dilemmas in a general theory of planning", Policy Sciences, 4(2), 1973, pp. 155–169.

[42] Seeber, I., G.-J. de Vreede, R. Maier, and B. Weber, "Beyond Brainstorming: Exploring Convergence in Teams", Journal of Management Information Systems, 34(4), 2017, pp. 939–969.

[43] Smith, M.L., K. Reilly, and Y. Benkler, "Open development: Networked innovations in international development", MIT Press, 2014.

[44] Su, B., A. Heshmati, Y. Geng, and X. Yu, "A review of the circular economy in China: Moving from rhetoric to implementation", Journal of Cleaner Production, 42, 2013, pp. 215–227.

[45] Sun, J. and K. Yang, "The Wicked Problem of Climate Change: A New Approach Based on Social Mess and Fragmentation", Sustainability, 8(12), 2016, p. 1312.

[46] Howe, J. The Rise of Crowdsourcing. http://www.wired.com/wired/archive/14.06/crowds.html.

[47] Thompson, M., "Ict and development studies: Towards development 2.0", Journal of International Development, 20(6), 2008, pp. 821–835.

[48] Tukker, A., M. Charter, C. Vezzoli, E. Stø, and M.M. Andersen, System Innovation for Sustainability 1: Perspectives on Radical Changes to Sustainable Consumption and Production, Taylor & Francis, 2017.

[49] Villarrubia-Gómez, P., S.E. Cornell, and J. Fabres, "Marine plastic pollution as a planetary boundary threat – The drifting piece in the sustainability puzzle", Marine Policy, 2017.

[50] Webster, K., The circular economy - a wealth of flows, 1st edn., Ellen MacArthur Foundation, Cowes, 2015.

[51] Wei, X. and W. Croft, "Investigating retrieval performance with manually-built topic models", Proceedings of Recherche d'Information Assistee par Ordinateur (RIAO), 2007, pp. 333–349.

[52] Weiss, S.M., N. Indurkhya, T. Zhang, and F.J. Damerau, Text Mining, Springer, 2005.

[53] World Economic Forum and Ellen MacArthur Foundation, The New Plastic Economy: Catalysing action, January 2017.

[54] World Economic Forum, Ellen MacArthur Foundation, and McKinsey & Company, The New Plastic Economy: Rethinking the future of plastics, January 2016.